General, PDB-based collective variables for protein folding

Ardevol, A., Palazzesi, F., Tribello, G. A., & Parrinello, M. (2016). General, PDB-based collective variables for protein folding. *Journal of chemical theory and computation*, *12*(1), 29-35. https://doi.org/10.1021/acs.jctc.5b00714

Published in:
Journal of chemical theory and computation

Document Version:
Peer reviewed version

General, PDB-based collective variables for protein folding

Albert Ardevol,*,† Ferruccio Palazzesi,† Gareth A. Tribello,‡ and Michele Parrinello†

Computational Science, Department of Chemistry and Applied Biosciences, ETH Zurich, USI-Campus, Via Giuseppe Buffi 13, C-6900 Lugano, Switzerland, and Atomistic Simulation Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast, BT7 1NN, United Kingdom

E-mail: albert.ardevol@phys.chem.ethz.ch

Abstract

New, automated forms of data-analysis are required in order to understand the high-dimensional trajectories that are obtained from molecular dynamics simulations on proteins. Dimensionality reduction algorithms are particularly appealing in this regard as they allow one to construct unbiased, low-dimensional representations of the trajectory using only the information encoded in the trajectory. The downside of this approach is that a different set of coordinates are required for each different chemical systems under study precisely because the coordinates are constructed using information from the trajectory. In this paper we show how one can resolve this problem by using the sketch-map algorithm that we recently proposed to construct a low-dimensional representation of the structures contained in the protein data bank (PDB). We show that the resulting coordinates are as useful for analysing trajectory data as coordinates constructed using landmark configurations taken from the trajectory and that these coordinates can thus be used for understanding protein folding across a range of systems.

*To whom correspondence should be addressed

†Computational Science, Department of Chemistry and Applied Biosciences, ETH Zurich, USI-Campus, Via Giuseppe Buffi 13, C-6900 Lugano, Switzerland

‡Atomistic Simulation Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast, BT7 1NN, United Kingdom

1 Introduction

For many years structural biologists have rationalised the functionality of proteins in terms of their tertiary structures.[1] As such an important question in the protein folding community has concerned the direct prediction of the tertiary (folded) structure from the primary amino acid sequence. Numerous groups have attempted to predict these structures based on a fundamental understanding of the interactions between the various atoms that make up the protein. However, as the number of protein structures determined by experimentalists has built up, an alternative approach based on homology between amino acid sequences has become popular. The plain fact is that the folded state of many proteins is generally some mixture of a relatively small number of secondary structural units of which the alpha helix and beta sheet are the two most common varieties. As a result one can make predictions about the tertiary structure of a new protein by looking for similar sequences of amino acids in the protein data bank (PDB). This new sequences will, in all probability, have a tertiary structure similar to those found in the PDB databank.

There is a growing consensus in the experimental community that the folded state is not the only factor that affects protein function. There is evidence that understanding phenomena such as allosteric binding,[2] signalling[3] or the growing class of so-called intrinsically disordered proteins[4] requires dynamical information on protein motions as well as static information on the structure of the folded state. It would be very difficult to extract such information using homology modelling alone so atomistic simulation based on an understanding of the individual interactions between atoms still has a clear role to play. In addition, it is still difficult to determine structures in loop regions of the protein using homology modelling.[5–7] Therefore, for all these problems molecular dynamics (MD) and Monte Carlo are thus still the methods of choice. Having said that homology modelling can play an important role when it comes to interpreting the results from such simulations or when developing variables to enhance the rate at which configuration space is sampled. Pietrucci and Laio[8] have shown that coordinates that count the number of segments of protein backbone that resemble the known α-helical and β-sheet secondary structure elements are useful collective variables. Furthermore, in many papers on atomistic simulation the results are rationalised using the lens provided by the known secondary structure elements.

In this paper we thus ask the question: can one develop a set of collective variables that can be used to insightfully interpret atomistic simulation data by using the information contained in the protein databank? To develop such coordinates we use the sketch-map algorithm[9] that we have recently developed. This dimensionality reduction algorithm has been shown to be remarkably robust[10] and in this paper we find that we can indeed interpret data generated during atomistic simulations using coordinates that are generated based on the known structures in the protein databank (PDB). The coordinates we are able to extract in this way allow us to differentiate between the various distinct configurations the protein adopts during the simulation. Furthermore, and perhaps more intriguingly, these coordinates tell us which configurations a particular amino acid sequence can adopt from amongst the constellation of possibilities that have been observed previously as well as those which it does not adopt.

2 Background

The sketch-map algorithm works in a manner similar to the classical multidimensional scaling (MDS) algorithm.[11] These dimensionality reduction algorithms endeavour to arrange a set of projections, $\{\mathbf{s}\}$, in some low dimensional space so that the distances between them are the same as the dissimilarities between the high-dimensional frames, $\{\mathbf{X}\}$, that the projections are supposed to represent. There are a number of non-linear dimensionality reduction (NLDR) algorithms[12–16] that are now used almost routinely to generate low-dimensional projections of data in this way, which all make assumptions about the structure of the high-dimensional data. In developing sketch-map we recognised that it is impossible to match all the dissimilarities between the trajectory frames at once as there are features in trajectory data that appear to be truly high dimensional. We thus chose to match a subset of dissimilarities only - the dissimilarities that are within a certain, user-specified range. Frames that are very similar are assumed to be essentially the same and are thus projected at the same point. Meanwhile the algorithm attempts to separate points that are very dissimilar but does not particularly worry about making the distance between the projections larger than the actual dissimilarity between the high-dimensional points.

In practise projections are generated in sketch-map by minimising the following function:

$$\begin{aligned} \chi^2 &= \sum_{i \neq j} w_i w_j \left[F(R_{ij}) - f(r_{ij})\right]^2 \\ \text{where} \qquad f(r) &= 1 - (1 + (2^{a/b} - 1)(r/\sigma)^a)^{-b/a} \\ \text{and} \qquad F(R) &= 1 - (1 + (2^{A/B} - 1)(R/\sigma)^A)^{-B/A} \end{aligned} \tag{1}$$

R_{ij} is the dissimilarity between landmark points $\mathbf{X}_i$ and $\mathbf{X}_j$ and r_{ij} is the distance between their projections $\mathbf{s}_i$ and $\mathbf{s}_j$. w_i and w_j are weights that are given to each of the landmarks. We generally calculate these weights by considering how many configurations from the trajectory lie in the Voronoi polyhedras of each of the landmarks. Within sketch-map the tuning to a distance range of interest is achieved by adjusting the value of σ. More information on this procedure as well as instructions for setting the other parameters in the above functions (A, B, a and b) can be found in the appendices of our recent paper[10] and on our website http://epfl-cosmo.github.io/sketchmap/.

When using sketch-map to examine trajectory data for protein molecules we use the values of the full set of Ramachandran angles to represent each protein configuration. As such when we use sketch-map to examine a 16 residue protein say each of the $\mathbf{X}$ vectors in our set of high dimensional configurations is a 30-dimensional vector of angles. To calculate the dissimilarity, R_{ij}, between two of these high-dimensional configurations we measure the square root of the sum of the squares of the differences between these backbone torsional angles. Furthermore, when calculating the difference between a pair of torsional angles we obviously take the fact that these quantities are periodic into account.

We have shown in recent papers[9,17] that, once projections for a relatively small number of landmark frames have been found, a projection, $\mathbf{x}$, for any high-dimensional configuration,

$\mathbf{X}$, can be found by minimizing:

$$\delta^2(\mathbf{x}) = \sum_{i=1}^{N} w_i \left\{ F[R_i(\mathbf{X})] - f[r_i(\mathbf{x})] \right\}^2 \tag{2}$$

where $R_i(\mathbf{X})$ is the dissimilarity between $\mathbf{X}$ and the ith landmark point and $r_i(\mathbf{x})$ is the distance between its projection, $\mathbf{x}$, and the projection of the ith landmark point. This procedure is remarkably robust. In fact we have shown that sketch-map coordinates generated from an MD trajectory on one particular chemical system can be used to examine a second chemical system even when the configurations adopted by this second system bear almost zero resemblance to those adopted by the first.[10] This formula is thus the basis of the work presented in this paper and our general coordinates for protein folding. In essence our approach involves using a small set of configurations from the PDB data bank and constructing projections for them using 1. We then insert these configurations and the projections we find for them into 2 and use this formula to generate projections of all the configurations visited during a molecular dynamics simulation. Once we have obtained a projection for each of the frames in our trajectory we can then construct a histogram, $H(\mathbf{x})$, (in this work by reweighting our biased trajectories using the method described in[18]). We then examine the structures that are projected near to the various maxima in this histogram as these configurations will have low free energies.

3 PDB-based coordinates

As discussed in the previous two sections in generating the sketch-map coordinates that have been employed in this work we took a different route from the one that we have adopted in previous papers. Instead of selecting a set of landmark coordinates from the MD trajectories we chose landmarks from the set of protein structures that have been determined by NMR and that have been deposited in the PDB databank. We chose to only use those structures that had been determined by NMR because this technique is usually applied to determine the structures of proteins with flexible domains in solution. We thus felt this would give us a wide enough range of peptide configurations to construct coordinates from.

For the first example, 1000 landmark points were selected from the set of configurations found in every 16-residue fragment contained in the 7846 NMR-solved structures deposited in the Protein Data Bank. In other words, the landmarks were selected from a library of more than 650,000 configurations that were generated by cutting all the NMR structures into 16-residue long segments. Clearly, these landmark configurations give a good representation of the full spectrum of conformational possibilities that have been observed for any 16-residue sequence of amino acids. The staged algorithm,[10] with a γ value of 0.1, was used to select a final set of landmark points, while the weights that enter equation 1 were determined by counting the fraction of the remaining PDB configurations that were in the Voronoi polyhedron of each of the landmarks. A sketch-map projection of this data was then generated by minimising equation 1 with $\sigma = 6.9$, $A = 8$, $B = 8$, $a = 1$ and $b = 4$. The resulting set of sketch-map coordinates were then used to analyze the conformational ensemble of the C-terminal fragment of the immunoglobulin binding domain B1 of protein G of

Streptococcus[19–22] (amino acids sequence Ace-GEWTYDDATKTFTVTE-NMe). We have recently performed extensive parallel tempering well tempered ensemble metadynamics (PT-WTE) simulations on this protein[23] and it was these trajectories that were analysed again in this work. In our previous work PT-WTE was used to enhance the sampling[24] as it has been shown to ensure fast and exhaustive exploration of the free-energy landscape.[22,25,26] These PT-WTE trajectories were generated using gromacs-4.5.5,[27] while torsional angles were calculated using PLUMED.[28,29] Further computational details for these calculations are presented in our recent article on the free energy landscape of this particular protein.[23]

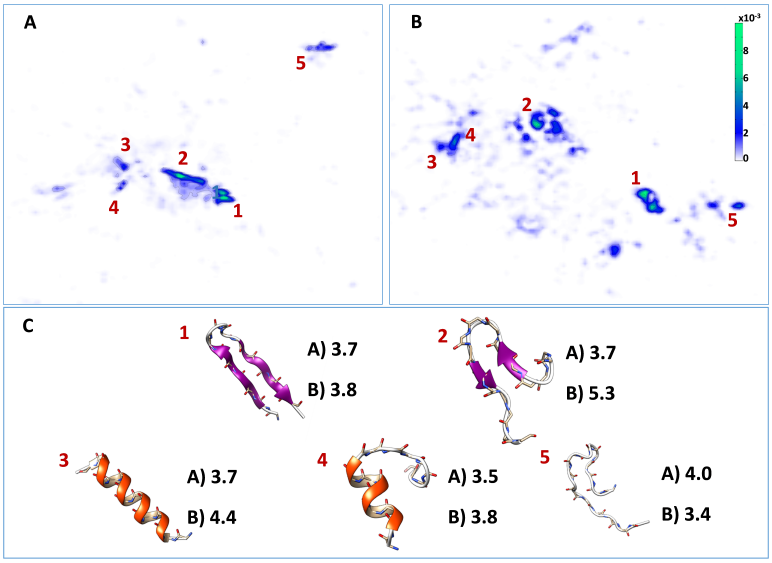

Figure 1: Histograms constructed from the trajectories of the β-hairpin protein studied in this work displayed as a function of the two sets of sketch-map coordinates. The histogram in panel A, H(WTE), is shown as a function of a set of sketch-map coordinates that were constructed using landmark configurations that were taken from the parallel tempering trajectories. Meanwhile, the histogram labelled B, H(PDB), is shown as a function of a set of sketch-map coordinates that were constructed using landmark configurations that were taken from the PDB data bank. Numerical labels were given to each of the various maxima in these histograms. Panel C shows representative configurations from each of these free energetic basins. The black numbers beside each of these protein configurations in panel C is the average root mean square deviation (r.m.s.d.) between the structures projected close to each of the maxima in the histogram. These quantities were calculated by taking the sum of the squares of the difference in torsional angles between configurations. The numbers labeled A and B are the average r.m.s.d. values calculated for H(WTE) and H(PDB) respectively.

To understand the free energy landscape that is being explored by this particular protein and in order to test the general, PDB-based sketch-map coordinates we constructed two representations of our trajectory data using sketch-map. For the first of these, we selected landmark configurations from the trajectory itself. In total 1000 landmark points were selected using the staged algorithm[10] with γ equal 0.1 and weights determined by counting the fraction of the remaining configurations that were in the Voronoi polyhedron of each of the landmarks. Projections of these configurations were generated using equation 1 with

$\sigma = 6$, $A = 8$, $B = 8$, $a = 2$ and $b = 8$. Projections for the remainder of the trajectory were then generated by minimizing equation 2. The histogram shown in figure 1A, H(WTE), was then generated by reweighting the trajectory.[18]

The second histogram, H(PDB), shown in figure 1B, was also generated by projecting and reweighting the data from our beta hairpin trajectory using equation 2. However, in this second case the landmark configurations for which projections were constructed using equation 1 were taken from the PDB databank as described previously. H(PDB) looks markedly different to H(WTE) but as we will show in the following paragraphs we can construct a mapping between the maxima in the two histograms. Before we do that, however, it is worth noting that both H(WTE) and H(PDB) show that the free energy landscape for this particular protein is extremely rough and that this protein is quite flexible. It can adopt a wide range of different structures all of which contain the familiar secondary structure units (e.g. α-helices, β-sheets, etc.) to different degrees. These different structural possibilities are all projected in different parts of the sketch-map planes, so these coordinates can be used to determine how frequently the protein adopts particular configurations. There are five configurations of the protein which appear from these coordinates to be particularly stable. This is based on the fact that there is a large maximum in the histogram at these points and, consequently, an associated minimum in the underlying free energy landscape. Representative configurations from these maxima are shown in figure 1C. Given the structures found in these basins they can be thought of as folded, misfolded and unfolded states of the protein.

In a recent paper[10] we discussed how one could extract a quantitative measure of the quality of any sketch-map projection by calculating the sketch-map stress function (equation 1 with all weights set equal to one) for those configurations whose projections were found by minimising equation 2. Obviously, minimising equation 1 for the many thousands of non-landmark points in a trajectory is not feasible but a one-time calculation of this quantity is not impossible. When this calculation is performed using the trajectory-based sketch map coordinates a value of 0.016 is obtained. By contrast, when the calculation is performed using the sketch-map coordinates that were constructed using the information from the PDB a value of 0.052 is obtained, which is slightly higher. However, as we discussed in our previous paper,[10] the quantity being calculated here measures how well the dissimilarities between non-landmark configurations are being reproduced in the two dimensional projections. These dissimilarities do not enter in equation 1 or equation 2 so this is a rather stringent test on the quality of our fitting procedure. Furthermore, the value of this quantity measures the fraction of configurations that are projected far apart when they should be close together or vice versa. 0.052 and 0.016 are thus both rather small fractions and hence this small decrease in quality does not particularly concern us.

An instructive exercise is to look at the structures that have their projections near the various maxima that we see in the histograms in figure 1. In order to facilitate this process we wrote a tcl package, called GISMO, that can be added the the popular molecular visualization pacakge VMD.[30] This package plots a small square in a tcl/tk canvas widget for each of the frames in the trajectory. These squares are centered on the values of a set of user-specified variables - in this paper we used the projections of the frames that were generated by sketch-map. The benefit of plotting using this tool is that the user can click on the squares and the corresponding trajectory frame is shown in the main VMD window. By using this tool

one is thus able to get a better understanding of where sketch-map is projecting the various configurations of the protein. This tool is thus extremely helpful when it comes to preparing figures such as that shown in figure 1 as we can use it to find representative configurations for each of the basins in the free energy landscape.

An examination of the projections and histograms generated with the two different sets of collective variables using the GISMO tool described in the previous paragraph shows us that, although the maxima appear in different locations in the two histograms, one can construct a one-to-one mapping between the maxima in H(PDB) and H(WTE). In other words, when you project the trajectory using any set of sketch-map coordinates you find a set of maxima in the histogram. Importantly, the set of high-dimensional structures to which these features correspond appear to be the same both when the sketch-map coordinates are constructed using trajectory data and when the coordinates are constructed using landmarks taken from the PDB. This gives us some confidence that we can identify the structures that have low-free energy by analysing the histogram of visited configuration displayed as a function of these PDB-based, sketch-map-generated coordinates.

To make the comparison between H(PDB) and H(WTE) more quantitative we collected all the structures that were projected in the vicinity of each of the basins highlighted in H(WTE) and performed some analysis on these structures. The first experiment we did involved calculating the dissimilarity between all of the structures within each of the basins separately. These dissimilarities were calculated as the square root of the sum of the squares of the differences in torsional angles in the different structures. In other words, we calculated these dissimilarities in the same manner that we calculated the dissimilarities between landmarks in equations 1 and 2. The average dissimilarity between the configurations in each of the basins is displayed close to the inset figures of the protein configurations. The number on top represents the average in-basin dissimilarity for the maxima corresponding to that structure in the left hand histogram H(WTE), while the lower number gives the same quantity for the histogram shown on the right H(PDB). The numbers obtained from this analysis are reassuring and suggest that all of the structures in each of the basins are structurally similar. For the majority of basins the average in-basin dissimilarity value is slightly higher for H(PDB), which suggests that these coordinates are slightly less discriminating than those constructed from trajectory data. This is perhaps to be expected, however, as the landmark configurations that were taken from the PDB cover a much broader range of structural possibilities than those taken from the trajectory. More importantly, however, this slight decrease in quality is does not large enough to put one off using the PDB-based sketch-map. The fact that it covers a broader range of structural possibilities makes it more general and hence more transferable. As such it can be used to project trajectories obtained for different chemical conditions or for different amino-acid sequences.

The results from the second experiment that we did to quantitatively compare the histograms obtained with the two sets of landmark configurations are shown in figure 2. To construct this figure we took the structures that were projected near to each of the maxima in H(WTE), and projected them using the sketch-map coordinates that were used to construct H(PDB). There are a very large number of frames projected near to the various maxima in H(WTE) therefore, to display this information more clearly, we built a histogram that shows where there these configurations are projected in H(PDB). These five histograms - one for each major basin - are shown as red contour plots in figure 2 overlain on H(PDB). It

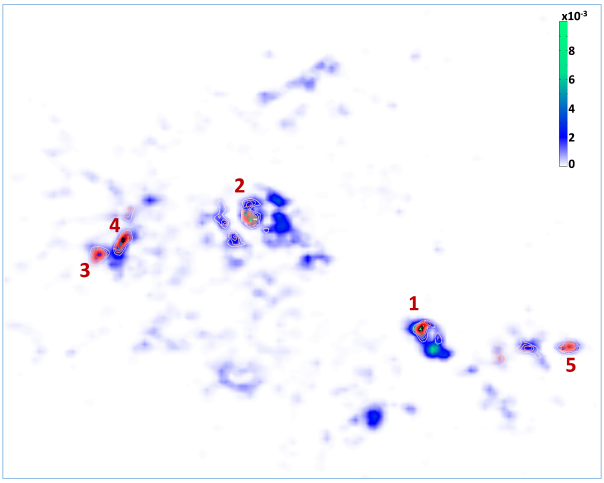

Figure 2: Figure showing the one to one mapping between the maxima in H(PDB) and H(WTE). The structures that were projected near to the five maxima in figure 1A. were collected using our in-house tool GISMO. These structures were then re-projected using the sketch-map coordinates that were built using landmark configurations taken from the PDB. A histogram showing the locations of the projections of these structures was then constructed. In the figure above this histogram is shown as a red contour plot on top of H(PDB). The same labels as were used in figure 1 are used to indicate the major basins. What is clear from this figures is that points that are projected close together when the data is projected using sketch-map coordinates constructed using trajectory data are also projected close together when they are projected using coordinates constructed using structures taken from the PDB. As a consequence we see that there is a one-to-one mapping between the low-free energy features that are seen in the left and right panels of figure 1.

is immediately apparent that those configurations that appear close together when they are projected using sketch-map coordinates that are constructed using landmarks taken from the trajectory, also appear close together when they are projected using the coordinates constructed using landmark data taken from the PDB. In addition, each of these markedly different sets of configurations are projected in different parts of the sketch-map plane (i.e. in different maxima) both when they are projected using coordinates constructed using trajectory data and when they are projected using coordinates constructed using data from the PDB. In addition, because we have displayed the histograms constructed by this procedure on top of H(PBE), we can see clearly that there is a one-to-one mapping between the maxima in H(WTE) and H(PDB). We are thus confident we can identify low free energy configurations using sketch-map and that the basins we observe are not simply artefacts from the dimensionality reduction's fitting procedure.

4 Understanding forcefield differences

Recently Palazzesi *et al.*[25] have studied a nine-residue intrinsically disordered peptide (IDP) whose dynamical behaviour has been experimentally characterized by Dames *et al.*[31] In their paper they show that the behavior of the peptide changed markedly when the interatomic potential was changed. In particular the distribution of end to end distances for the AMBER03w[32,33] force field is shifted to longer values with respect to that observed with AMBER99SB*-ILDN[34–36] as shown in the inset in figure 3. These marked differences between the behaviours of the two force fields were also observed for a number of biophysical parameters and for some NMR experimental observables.[25] Palazzesi *et al.* were thus eventually able to conclude that the AMBER03w was the better forcefield for this particular IDP as this potential reproduced the experimentally observed C^{α} and N^{H} chemical shifts and $^{3}J(H^{N}\text{-}H^{\alpha})$ coupling parameters better than the AMBER99SB*-ILDN.

Differences in the ensemble averages for two force fields come about because each force field stabilizes a different set of atomic configurations. Consequently, a better understanding of the conformational ensemble that is being explored with a particular force field allows one to explain more completely why averaged properties, such as the NMR observables, have the value that they do. In this regard the two dimensional projections generated by sketch-map are useful as they help you to examine the conformational ensemble. In other words, we can use sketch-map to project the trajectories obtained with the AMBER03w and AMBER99SB*-ILDN force fields and can thus visualize the differences in the ensembles of configurations visited during the two trajectories. This procedure gives one more information on the structures that are being adopted than simply comparing the distribution of end-to-end distances or the distribution of values observed for some other biophysical characteristic. However, a difficulty with this approach would have been deciding whether to use a sketch-map projection in which the landmark frames were selected from the AMBER03w trajectories, a sketch-map projection constructed from the AMBER99SB*-ILDN trajectory or a sketch-map projection that contained landmarks from both trajectories. With our new approach based on using the protein data bank, however, we no longer need to make this decision. We constructed sketch-map coordinates for this system using landmarks taken from the PDB. Once again we downloaded the 7,846 structures that have been determined

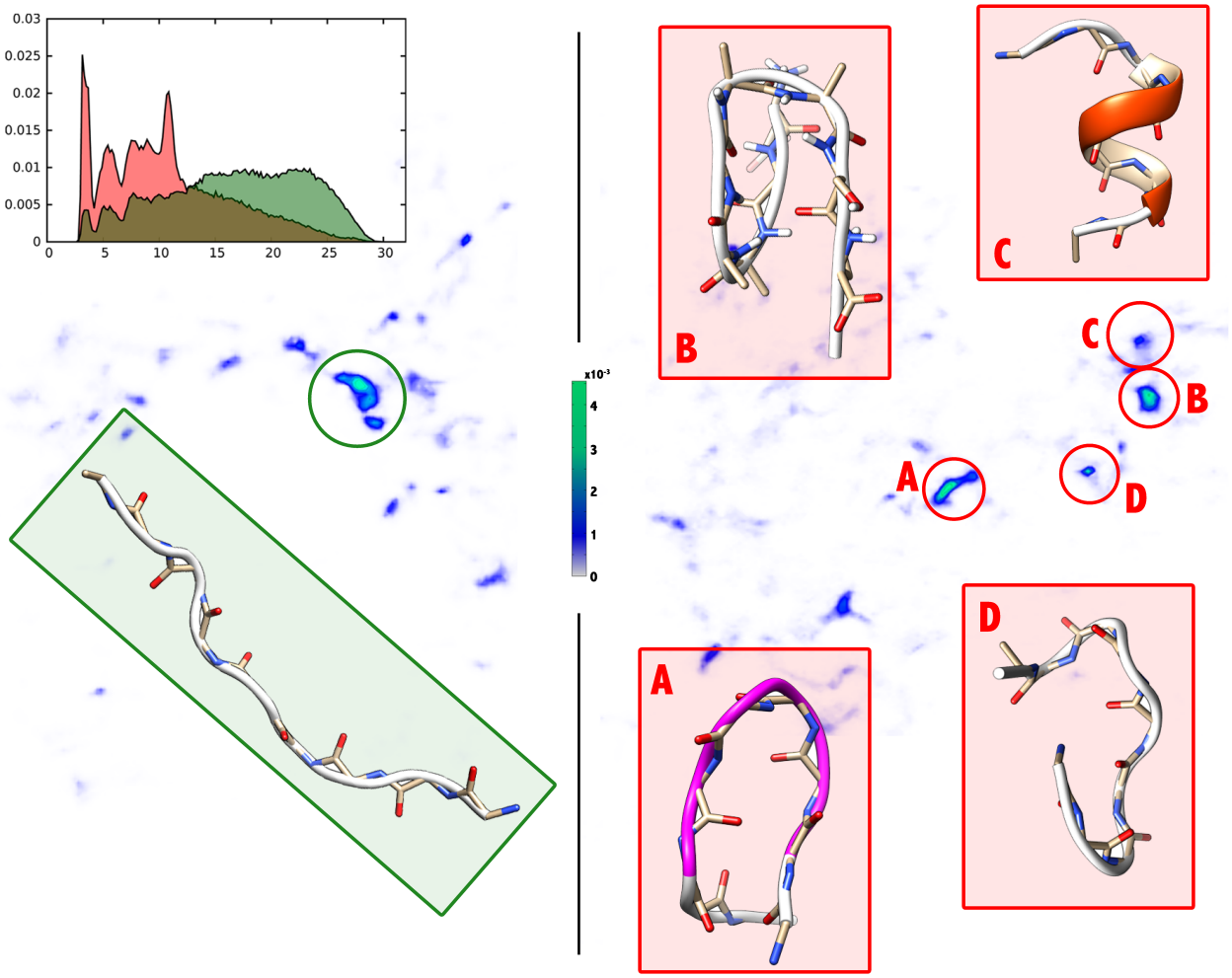

Figure 3: Histograms of the trajectory for the 9-residue, intrinsically disordered protein studied in Palazzesi's work shown as a function of the general PDB-based sketch-map coordinates. The histogram on the left was calculated using a trajectory where the AMBER03w forcefield was used to evaluate the forces while that on the right was calculated using the AMBER99SB*-ILDN forcefield. The structures corresponding to the most populated basins are shown in the insets. The AMBER03w forcefield has only one prominent basin in its landscape and the configurations in this basin all have the chain fully extended. Meanwhile, the AMBER99SB*-ILDN forcefield has numerous basins in its free energy landscape for which the protein is compact and contains some secondary structural elements. These differences between the configurations adopted by the two forcefields explain why the distribution of end to end distances is peaked at a higher value for the AMBER03w forcefield. The distribution of end-to-end distances for the two forcefields is shown in the inset. The green curve is the distribution for the AMBER03w forcefield, while the red curve is the distribution for the AMBER99SB*-ILDN.

using NMR but this time we cut them into 9 residue segments. This procedure gives more than 700,000 structures from which we selected 1,000 landmark frames using the staged algorithm as described in the previous sections. Once again we determined the weights that enter equation 1 by considering how many of the non-landmark points from the PDB are in the Voronoi polyhedra of each of the landmarks. Projections of these landmark frames were then constructed by minimizing 1 with $\sigma = 4.2$, $A = 7$, $B = 7$, $a = 1$ and $b = 4$. The original frames and these projections were then inserted into equation 2, which was used to project the parallel-tempering well-tempered ensemble (PT-WTE)[37,38] trajectories generated in Palazzesi *et al.*'s paper. The resulting histograms for this 9-residue protein, modelled with the two different force fields, and projected on this general set of sketch-map coordinates are shown in figure 3.

Figure 3 shows that the configurations the protein adopts in the simulations with the two force fields are markedly different. When the atomic interactions are modelled using the AMBER99SB*-ILDN potential the system samples from the four high maxima in the histogram shown around the right panel of 3. There is evidence of secondary structural elements in the configurations projected near these features. Conformations A and C, in particular, resemble an α-helix and a β-sheet respectively.

The histogram obtained when the calculations are run using the AMBER03w potential shows only one marked maxima. The protein configurations that are projected near to this maxima are extended and show no evidence of any particular secondary structural elements. The fact that this is the most stable configuration of the peptide for AMBER03w explains why the average end-to-end distance is very long in this simulation - this trajectory contains very few structured configurations. By contrast, the average end-to-end distances from the AMBER99SB*-ILDN simulations are considerably shorter in large part because when this potential is used the system remains for long periods of time in collapsed, secondary-structure-containing configurations.

5 Conclusion

Our previous papers[9,10,17] have demonstrated that sketch-map is a useful tool for analysing the output from molecular dynamics simulations. By projecting representative configurations from the trajectory into a lower dimensional space an unbiased set of collective coordinates is created. These collective coordinates are less reliant on a simulators chemical or physical intuition about the problem, which increases the likelihood of making surprising or unexpected discoveries. In previous work we have always constructed sketch-map variables by selecting landmark frames from the molecular dynamics trajectories we have run. In this work we have shown how we can construct sequence-independent coordinates using experimental data and that we can use these coordinates to interpret trajectories, even if they sample very different parts of conformational space. The use of landmarks from the protein data bank ensures that the coordinates contain a reasonable representation of the various structural possibilities that are open to the peptide. Furthermore, by not selecting landmarks from the trajectory one ensures that the quality of the sketch-map coordinates will not be affected if the trajectory does sample all of configuration space. Lastly, standard general PDB-based sketch-map coordinates can be constructed for amino acid sequences of different

lengths and placed online in a repository. Other researchers using sketch-map can then profit by using these variables and comparing the results that they obtain with those that were obtained in prior publications or in different simulation conditions. We have thus provided various sets of sketch-map coordinates for peptides ranging from six to sixteen amino acids in length. These coordinate sets are available from http://epfl-cosmo.github.io/sketchmap/.

Sketch-map is not solely for the analysis of trajectory data. Elsewhere we have demonstrated that bias potentials can be constructed that are a function of the sketch-map coordinates and that the sampling of phase space can be accelerated in this way.[17] We have shown that the advantage of this over more conventional approaches is that the sketch-map variables are able to distinguish between many of the basins in the free energy landscape. As such when these variables are used in tandem with methods such as metadynamics fewer problems arise because of barriers in the transverse degrees of freedom. In this context the general, PDB-based sketch-map coordinates that have been the subject of this paper represent an exciting possibility. If such coordinates can be used to enhance sampling this will greatly reduce the amount of time that has to be spent looking for appropriate collective variables to study protein folding. Using the information contained in the protein data bank ensures that the structural possibilities for the amino acid sequence are enumerated in the coordinates. Furthermore, because of the way sketch-map projections are constructed, different basins in the energy landscape will be projected in different parts of the low dimensional space. These PDB-based sketch-map variables thus have the potential to be general collective variable for metadynamics simulations, which is an idea that we intend to explore further in future publications.

6 Acknowledgements

The authors would like to thank Alessandro Barducci and Michele Ceriotti for useful discussions. Calculations were performed using the Rosa supercomputer at the Swiss National Super Computer Centre (CSCS) under project ID s223. A.A. was supported by an EMBO long-term fellowship. Molecular graphics and analyses for the insets in the figures in this paper were generated using the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualisation and informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).[39]

7 Supporting Information

The Supporting Information is available free of charge via the Internet at http://pubs.acs.org Guidance on setting the Sketch-map parameters and some general comments on the applicability of the Sketch-map algorithm.

References

(1) Lodish, H.; Berk, A.; Matsudaira, P.; Kaiser, C. A.; Krieger, M.; Scott, M. P.; Zipursky, S. L.; Darnell, J. *Molecular Cell Biology*, Fifth ed.; W H Freeman: New

York, 2003.

(2) Christopoulos, A. *Nature Reviews Drug Discovery* **2002**, *1*, 198–210.

(3) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. *Current Opinion in Structural Biology* **2008**, *18*, 756–764.

(4) Dyson, H. J.; Wright, P. E. *Nature Reviews Molecular Cell Biology* **2005**, *6*, 197–208.

(5) Constanzi, S. *Chim. Oggi.* **2010**, *28*, 26–31.

(6) Goldfeld, D. A.; Zhu, K.; Beuming, T.; Friesner, R. A. *Proceedings of the National Academy of Sciences* **2011**, *108*, 8275–8280.

(7) Kmiecik, S.; Jamroz, M.; Kolinski, M. *Biophysical Journal* **2015**, *106*, 2408 – 2416.

(8) Pietrucci, F.; Laio, A. *Journal of Chemical Theory and Computation* **2009**, *5*, 2197–2201.

(9) Ceriotti, M.; Tribello, G. A.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13023–13029.

(10) Ceriotti, M.; Tribello, G. A.; Parrinello, M. *Journal of Chemical Theory and Computation* **2013**, *9*, 1521–1532.

(11) Borg, I.; Groenen, P. J. *Modern Multidimensional Scaling*, 2nd ed.; Springer: New York, 2005.

(12) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. *Science* **2000**, *290*, 2319–2323.

(13) Roweis, S. T.; Saul, L. K. *Science* **2000**, *290*, 2323–2326.

(14) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. *Proc. Natl. Acad. Sci. USA of the United States of America* **2005**, *102*, 7432–7437.

(15) Coifman, R. R.; Lafon, S. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5 – 30.

(16) Belkin, M.; Niyogi, P. *Neural Comput.* **2003**, *15*, 1373–1396.

(17) Tribello, G. A.; Ceriotti, M.; Parrinello, M. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 5196–5201.

(18) Bonomi, M.; Barducci, A.; Parrinello, M. *J. Comput. Chem.* **2009**, *30*, 1615–1621.

(19) Blanco, F. J.; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584–590.

(20) Hughes, R.; Waters, M. *Curr. Opin. Struct. Biol.* **2006**, *16*, 514–524.

(21) Muñoz, V.; Thompson, P.; Hofrichter, J.; Easton, W. *Nature* **1997**, *390*, 196–199.

(22) Deighan, M.; Bonomi, M.; Pfaendtner, J. *Journal of Chemical Theory and Computation* **2012**, *8*, 2189–2192.

(23) Ardevol, A.; Tribello, G. A.; Ceriotti, M.; Parrinello, M. *Journal of Chemical Theory and Computation* **2015**, *11*, 1086–1093, PMID: 26579758.

(24) Barducci, A.; Bussi, G.; Parrinello, M. *Physical review letters* **2008**, *100*, 020603.

(25) Palazzesi, F.; Barducci, A.; Tollinger, M.; Parrinello, M. *Proceedings of the National Academy of Sciences* **2013**, *110*, 14237–14242.

(26) Barducci, A.; Bonomi, M.; Prakash, M. K.; Parrinello, M. *Proceedings of the National Academy of Sciences* **2013**, *110*, E4708–E4713.

(27) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(28) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. *Comp. Phys. Comm.* **2009**, *180*, 1961 – 1972.

(29) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Computer Physics Communications* **2014**, *185*, 604 – 613.

(30) Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(31) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *Journal of the American Chemical Society* **2006**, *128*, 13508–13514.

(32) Best, R. B.; Mittal, J. *The Journal of Physical Chemistry B* **2010**, *114*, 14916–14923.

(33) Abascal, J. L.; Vega, C. *The Journal of chemical physics* **2005**, *123*, 234505.

(34) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 712–725.

(35) Best, R. B.; Hummer, G. *The Journal of Physical Chemistry B* **2009**, *113*, 9004–9015.

(36) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Structure, Function, and Bioinformatics* **2010**, *78*, 1950–1958.

(37) Earl, D. J.; Deem, M. W. *Physical Chemistry Chemical Physics* **2005**, *7*, 3910–3916.

(38) Bonomi, M.; Parrinello, M. *Physical Review Letters* **2010**, *104*, 190601.

(39) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *Journal of Computational Chemistry* **2004**, *25*, 1605–1612.

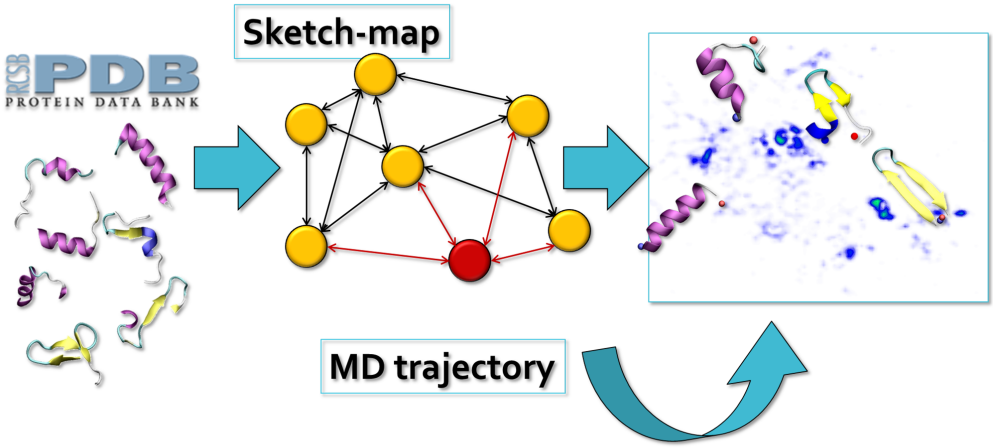

Figure 4: For Table of Contents Only (TOC)